Steady-state distributions for human decisions in two-alternative choice tasks

Stewart, Andrew; Cao, Ming; Leonard, Naomi Ehrich

Published in:
Proceedings of the 2010 American Control Conference

Publication date:
2010

2010 American Control Conference
Marriott Waterfront, Baltimore, MD, USA
June 30-July 02, 2010

Steady-state distributions for human decisions in two-alternative choice tasks

Andrew Stewart, Ming Cao and Naomi Ehrich Leonard

Abstract— **In human-in-the-loop systems, humans are often faced with making repeated choices among finite alternatives in response to observations of the evolving system performance. In order to design humans into such systems, it is important to develop a systematic description of human decision making in this context. We examine a commonly used, drift-diffusion, decision-making model that has been fit to human neural and behavioral data in sequential, two-alternative, forced-choice tasks. We show how this model and type of task together can be regarded as a Markov process, and we derive the steady-state probability distribution for choice sequences. Using the analytic expression for this distribution, we prove matching behavior for tasks that exhibit a matching point and we compute the sensitivity of steady-state choices to a model parameter that measures the decision maker's "exploratory" tendency.**

I. Introduction

It is not uncommon in human-in-the-loop systems that humans will be confronted repeatedly with decision-making problems in which, having observed the performance of the system, they must choose between two or more alternatives in order to maintain or improve performance. For example, in [1], the authors study human supervisory control of multiple unmanned aerial vehicles where choices must be made between attending to targets and ensuring safe return of vehicles. Human flight control operators face many such choices, for example, in bad weather when it must be decided whether or not to ground each of many vehicles [2]. The authors of [3] explore a setting in which a human must repeatedly choose one of two different robotic oxygen extraction systems operating on Mars with the goal of maximizing long-term oxygen extraction; the investigation focuses on the well-known difficulty that humans have with making long-term optimal decisions when short-term performance is high.

A systematic description of human decision making can be of critical value in designing such human-in-the-loop systems. In this paper we focus on human decision making in tasks where each choice is to be made between two alternatives. *Two-alternative forced-choice* (TAFC) tasks have been used extensively in the psychology literature to investigate human decision-making behavior in decision-making problems that require sequential binary choices of this sort [4]. A number of studies have focused explicitly on the case in which a performance measure, referred to as a *reward*, is provided to the human subject after every choice, and the reward is a function not only of the immediate choice but also of the subject's recent history of choices [4], [5], [6]. The human subject can then base the next decision on the current (and past) rewards received. The dependence of performance on past decisions is highly relevant for real-world human-in-the-loop decision-making problems.

The successful fitting of both behavioral and neural data taken from TAFC task experiments provides strong justification for the widespread use of the *Drift Diffusion Model* (DDM) to describe human decision making in TAFC tasks [4], [6]. Further, the DDM can be derived from the dynamics of a variable that represents the evidence in neuronal populations in favor of one alternative over the other [7]. It can also be interpreted as a continuum limit of the Sequential Probability Ratio Test [7].

Motivated by the challenges in designing human-in-the-loop systems, we leverage the TAFC task research and in particular use the DDM to derive formal expressions for behavior and performance sensitivity for decision making in this context. We do this by proving that the model is Markov under two important simplifying assumptions, the stronger of which is used and justified in [4].

We describe the TAFC task in Section II and the DDM for decision making in Section III. We prove that the model is Markov in Section IV and derive the steady-state choice distribution in Section V. In Section VI we prove results on the steady-state decision-making dynamics. We make final remarks in Section VII.

II. Two-alternative forced-choice task

Montague and co-authors [4], [6] introduced the two-alternative forced-choice (TAFC) task in which the decision maker is required to make a choice between two alternatives (denoted A and B), sequentially in time, and a reward (performance measure) is received after each choice is made. The decision maker's goal is to maximize total accumulated reward (optimize performance over the long run). Figures 1 and 2 show example reward schedules that are used in behavioral studies; the reward r_A for choosing A (resp. r_B for B) is plotted as a dashed line (resp. solid) as a function of y, which is the fraction of times A is chosen in the past N decision trials.

Figure 1 is called the *matching shoulders* reward structure [6] and represents the case in which there are diminishing returns for choosing A for too long and likewise for choosing B for too long. The point at which the curves intersect is called the matching point, and there is extensive empirical

A. Stewart and N. E. Leonard are with Department of Mechanical and Aerospace Engineering, Princeton University, USA ({arstewar,naomi}@princeton.edu). M. Cao is with Faculty of Mathematics and Natural Sciences, ITM, University of Groningen, the Netherlands (m.cao@rug.nl).

This research was supported in part by AFOSR grant FA9550-07-1-0-0528 and ONR grant N00014–04–1–0534.

evidence that human decision makers converge in aggregate to choice sequences y that correspond to the matching point. This is despite the fact that the expected value of the reward at the matching point is not necessarily optimal as seen in Figure 1. Figure 2 is called the *converging gaussians* reward structure and has been used recently in empirical studies of decision dynamics in social TAFC tasks [8]. The converging gaussians reward structure also has a matching point and experiments show that decision makers converge to the matching point [8].

Let $x(t) = (x_1(t), x_2(t), \ldots, x_N(t))$ denote the past N choices ordered sequentially in time with $x_1(t) \in \{A, B\}$ the most recent decision at time t, $x_2(t) \in \{A, B\}$ the most recent decision at time $t-1$, etc. We have that

$$x_k(t+1) = x_{k-1}(t), \qquad k = 2, \ldots, N, \ t = 0, 1, 2, \ldots \quad (1)$$

Let $y(t)$ denote the proportion of choice A in the last N trials at time t; i.e.

$$y(t) = \frac{1}{N} \sum_{k=1}^{N} \delta_{kA}(t) \quad (2)$$

where

$$\delta_{kA}(t) = \begin{cases} 1 & \text{if } x_k(t) = A \\ 0 & \text{if } x_k(t) = B. \end{cases}$$

Note that y can only take values from a finite set $\mathcal{Y}$ of $N+1$ discrete values:

$$y \in \mathcal{Y} = \left\{ \frac{i}{N}, \ i = 0, 1, \ldots, N \right\}.$$

The reward at time t is given by

$$r(t) = \begin{cases} r_A(y(t)) & \text{if } x_1(t) = A \\ r_B(y(t)) & \text{if } x_1(t) = B. \end{cases} \quad (3)$$

We define the difference in the reward as

$$\Delta r(y(t)) := r_B(y(t)) - r_A(y(t)). \quad (4)$$

The dynamics of the human decision-making process in the TAFC task can be modeled as an N-dimensional, discrete-time dynamical system where $x(t)$ is the state of the system and $y(t)$ is the output of the system.

III. Drift Diffusion Model

The Drift Diffusion Model (DDM) for decision making derives from a one-dimensional drift diffusion process described by a stochastic differential equation [8], [9], [10]:

$$dz = \alpha dt + \sigma dW, \qquad z(0) = 0. \quad (5)$$

Here z represents the accumulated evidence in favor of a candidate choice of interest, α is the drift rate representing the signal intensity of the stimulus acting on z and σdW is a Wiener process with standard deviation σ, which is the diffusion rate representing the effect of white noise.

Now consider the TAFC task with choices A and B. The drift rate α, as described in [6], [11], is determined by a subject's anticipated rewards w_A and w_B for a decision of A or B. Take z to be the accumulated evidence for choice A relative to choice B. Then on each trial a choice is made

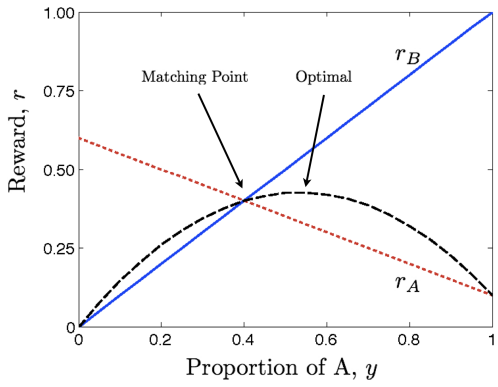

Fig. 1. The *matching shoulders* reward structure [4]. The dotted line depicts r_A, the reward for choice A. The solid line depicts r_B, the reward for choice B. The dashed line is the average value of the reward. Each is plotted against y, the fraction of choice A made in the last N trials.

when $z(t)$ *first* crosses one of the predetermined thresholds $\pm\nu$. If $+\nu$ is crossed then choice A is made, and if $-\nu$ is crossed then choice B is made. For such drift diffusion processes, as pointed out in [8], it can be computed using tools developed in [7] that the probability of choosing A in the next time step is

$$p_A(t+1) = \frac{1}{1 + e^{-\mu(w_A(t) - w_B(t))}} \quad (6)$$

where $\mu(w_A - w_B)$ is identified with $2(\alpha/\sigma)^2(\nu/\alpha)$. The right side of equation (6) is a sigmoidal function of $w_A - w_B$ where μ is the slope. Larger μ implies more certainty in the decision making, sometimes interpreted as less of a tendency to explore.

Studies of the role of dopamine neurons in coding for reward prediction error [12] have motivated the use of temporal difference learning theory [13] to describe the update of w_A and w_B. Let $Z \in \{A, B\}$ be the choice made at time t, then

$$w_Z(t+1) = (1-\lambda)\omega_Z(t) + \lambda r(t) \quad (7)$$

$$w_{\bar{Z}}(t+1) = w_{\bar{Z}}(t) \qquad t = 0, 1, 2, \ldots \quad (8)$$

where $\bar{\cdot}$ denotes the "not" operator. Here, $\lambda \in [0, 1]$ acts as a learning rate, controlling how the anticipated reward of choice Z at $t+1$ is affected by its value at t.

IV. Markov Model of Decision Making

Consider the DDM decision maker faced with the two-alternative, forced-choice task. As the DDM makes sequential decisions and receives corresponding rewards, the proportion of choice A evolves in time according to the dynamics of the coupled decision maker and task system described in Sections II and III. In this section we find conditions under which the decision making can be modeled as a Markov process. We derive the probability transition function for $y(t)$ and build a one-step transition matrix which is used in Section V to compute the steady-state distribution for the process.

The full state of the DDM in the two-alternative, forced-choice task is the N-element decision history $x(t)$, coupled with the expected rewards $w_A(t)$ and $w_B(t)$. To reduce the order of the system we make the following assumptions:

Assumption 1: $\Pr\{x_k(t) = A|x(t)\} = y(t)$

Assumption 2: $w_B(t) - w_A(t) = \Delta r(y(t))$.

Assumption 1 implies that the yN A's and $(1 - yN)$ B's in $x(t)$ are uniformly distributed in the finite history. Assumption 2 sets the difference in anticipated rewards at time t equal to the difference in rewards evaluated at $y(t)$; according to Montague and Berns [4] this assumption is true "on average". Together these assumptions reduce the dimension of the state space to one.

Proposition 1: Suppose Assumptions 1 and 2 hold. Then, the DDM (6) for the TAFC task (1)-(3) is a Markov Process with state $y(t)$ and transition probabilities given by

$$\Pr\{y(t+1) = y(t) - \frac{1}{N}\} = \frac{e^{\mu\Delta r}y(t)}{1 + e^{\mu\Delta r}} \tag{9}$$

$$\Pr\{y(t+1) = y(t)\} = \frac{e^{\mu\Delta r} + (1 - e^{\mu\Delta r})y(t)}{1 + e^{\mu\Delta r}} \tag{10}$$

$$\Pr\{y(t+1) = y(t) + \frac{1}{N}\} = \frac{1 - y(t)}{1 + e^{\mu\Delta r}} \tag{11}$$

where $\Delta r = \Delta r(y(t))$ is given by (4).

Proof of Proposition 1:

Since for a given choice $x_1(t+1)$ at time $t+1$, $y(t+1)$ can only change from its current value of $y(t)$ to $y(t) + \frac{1}{N}$, $y(t) - \frac{1}{N}$ or stay at $y(t)$, we need only compute the probability of each of these three events for all $y(t) \in \mathcal{Y}$. Each of these events depends upon the current value of $y(t)$ as well as $x_1(t+1)$ and $x_N(t)$ since $y(t+1)$ will only differ from $y(t)$ if $x_1(t+1)$ also differs from $x_N(t)$.

The event that $y(t+1) = y(t) - \frac{1}{N}$ requires $x_1(t+1) = B$ and $x_N(t) = A$. Treating these as independent events and using Equation (6) with Assumption 1 yields

$$\begin{aligned}\Pr\{y(t+1) = y(t) - \frac{1}{N}\} &= \\ &\Pr\{x_1(t+1) = B\} * \Pr\{x_N(t) = A\} \\ &= \frac{e^{\mu(w_B(t) - w_A(t))}y(t)}{1 + e^{\mu(w_B(t) - w_A(t))}}.\end{aligned}$$

Substituting in the identity of Assumption (2), we get Equation (9).

Similarly, the probability that $y(t+1)$ takes the value $y(t) + \frac{1}{N}$ is given by

$$\begin{aligned}\Pr\{y(t+1) = y(t) + \frac{1}{N}\} &= \\ &\Pr\{x_1(t+1) = A\} * \Pr\{x_N(t) = B\} \\ &= \frac{1 - y(t)}{1 + e^{\mu(w_B(t) - w_A(t))}}.\end{aligned}$$

Substituting in the identity of Assumption (2), we get Equation (11).

The event that $y(t+1) = y(t)$ requires either $x_1(t+1) = A$ and $x_N(t) = A$ or $x_1(t+1) = B$ and $x_N(t) = B$. The probability of the union of these events is

$$\begin{aligned}\Pr\{y(t+1) = y(t)\} &= \\ &\Pr\{x_1(t+1) = A\} * \Pr\{x_N(t) = A\} \\ &+ \Pr\{x_1(t+1) = B\} * \Pr\{x_N(t) = B\} \\ &= \frac{y(t) + (1 - y(t))e^{\mu(w_B(t) - w_A(t))}}{1 + e^{\mu(w_B(t) - w_A(t))}}.\end{aligned}$$

Substituting in the identity of Assumption (2), we get Equation (10). Since the probabilities depend on $y(t)$ only, the state at time t, the process is Markov. □

Equations (9)-(11) are used to build the $(N+1) \times (N+1)$ one-step transition matrix $\mathbf{P}$ which has entries $P_{ij} = \Pr\{y(t+1) = \frac{j}{N} | y(t) = \frac{i}{N}\}, i, j \in \{0, 1, \ldots, N+1\}$.

V. Steady-State Choice Distribution

Since the Markov process modeled in Section IV is irreducible and aperiodic, it has a unique limiting distribution $\pi = (\pi_0, \pi_1, \ldots, \pi_N)$ describing the fraction of time the chain will spend in each of the enumerated states in the long run (as $t \to \infty$) [14]. This steady-state distribution is the solution to the following equations:

$$\pi\mathbf{P} = \pi \tag{12}$$

$$\sum_{i=0}^{N} \pi_i = 1. \tag{13}$$

Proposition 2: For the transition probabilities given by (9) - (11) the unique steady-state distribution is

$$\pi_i = \frac{\alpha_i(1 + e^{\mu\Delta r(\frac{i}{N})})e^{-\mu\beta_i}}{\sum_{j=0}^{N} \alpha_j e^{-\mu\beta_j}(1 + e^{\mu\Delta r(\frac{j}{N})})} \tag{14}$$

where $\alpha_i = \frac{N!}{(N-i)!i!}$ and $\beta_i = \sum_{j=1}^{i} \Delta r(\frac{j}{N})$.

Proof of Proposition 2: Solving (12) alone yields a row vector v whose elements are given by

$$v_i = \frac{N!}{(N-i)!i!}(1 + e^{\mu\Delta r(\frac{i}{N})})e^{-\mu\sum_{j=1}^{i}\Delta r(\frac{j}{N})}.$$

To solve (13) we normalize the vector v to get

$$\pi = \frac{v}{\sum_{i=0}^{N} v_i}.$$

The elements of π are then given by Equation (14). □

Figure 2 shows the converging gaussians reward structure [8] along with the corresponding steady-state choice distribution π. Note that the distribution peaks where the reward curves intersect (the *matching point*); i.e. the decision maker spends the highest fraction of time with proportion of choice A at the matching point. This is in agreement with experimental results shown in Figure 3 of [8]. We show more general conditions under which this occurs in Section VI, where we also derive the sensitivity to μ of the reward earned by the decision maker.

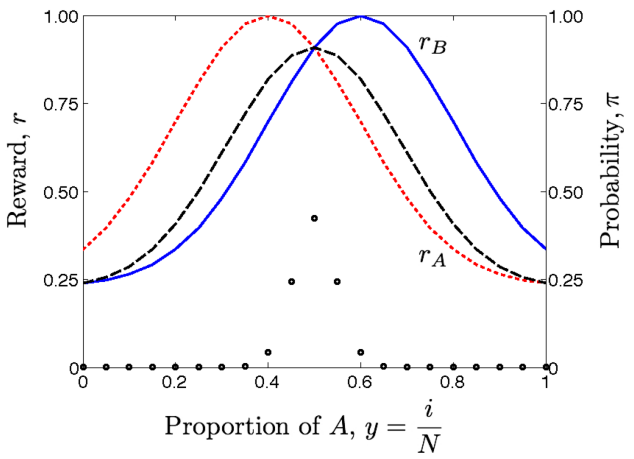

Fig. 2. The *converging gaussians* reward structure [8]: The dotted line depicts r_A, the reward for choice A. The solid line depicts r_B, the reward for choice B. The dashed line is the average value of the reward. The limiting distribution π is given by Equation (14) and is shown for $N = 20$ and $\mu = 8$ by the circular points. Each component π_i is plotted against $y = \frac{i}{N}$.

VI. Analysis of Steady-State Distribution

In this section we perform two analyses of the steady-state behavior. First we show conditions for which the decision maker converges to a value of y that corresponds to a matching point. To do this we consider *reward structures with a unique matching point* (Figures 1 and 2 are two examples of such reward structures). Second we derive the sensitivity of the expected value of reward earned by the decision maker to the parameter μ in the DDM (6).

A. Steady-State Matching

Matching behavior is a well-known phenomenon in human behavioral experiments [15], [16]: human decision makers in TAFC tasks converge in aggregate to choice sequences in the neighborhood of the matching point for a variety of reward structures that have a matching point. However, there are relatively few results that prove conditions for this phenomenon given well-established models like the DDM. In [4], Montague and Berns use Assumption 2 to show that the matching point in the matching shoulders reward structure is an attracting point. In [17], [18] a proof of convergence to a neighborhood of the matching point is shown for the Win Stay Lose Switch (WSLS) decision-making model and a deterministic limit of the DDM. A related analysis for the WSLS model is performed in [19].

In this section we prove steady-state matching behavior for the DDM by finding sufficient conditions on the slope μ of the DDM that guarantee that π_i is greatest for $y = i/N$ at or near the matching point. In Theorem 1 below, we find a bound μ_1 such that if $\mu > \mu_1$ then π_i peaks in a small neighborhood of the matching point. In Theorem 2 we find a bound $\mu_2 > \mu_1$ such that if $\mu > \mu_2$ then π_i peaks at the matching point.

Definition 1: Let a *reward structure with a single matching point* consist of reward curves $r_A(y)$, $r_B(y)$ for which there exists $y^* = \frac{i^*}{N}$, $i^* \in \{1, 2, \ldots, N-1\}$, that satisfies $\Delta r(y^*) = 0$, $\Delta r(y) < 0$ for $y < y^*$, and $\Delta r(y) > 0$ for $y > y^*$.

Theorem 1: If reward structures satisfy Definition 1 and

$$\mu > \mu_1 := \max\left\{ -\frac{\ln\left(\frac{\lfloor \frac{N}{2}\rfloor!\lceil \frac{N}{2}\rceil!}{(N-i^*)!i^*!}\right)}{\left|\Delta r\left(\frac{i^*+1}{N}\right)\right|}, -\frac{\ln\left(\frac{\lfloor \frac{N}{2}\rfloor!\lceil \frac{N}{2}\rceil!}{(N-i^*)!i^*!}\right)}{\left|\Delta r\left(\frac{i^*-1}{N}\right)\right|} \right\} \tag{15}$$

then the steady-state choice distribution is maximum for $y \in \{y^* - \frac{1}{N}, y^*, y^* + \frac{1}{N}\}$; where $\lfloor \cdot \rfloor$ gives the largest integer less than its argument and $\lceil \cdot \rceil$ gives the smallest integer greater than its argument.

Proof of Theorem 1: To prove Theorem 1 we examine $\rho(i) = \pi_i/\pi_{i^*}$, the ratio of time spent at $y = \frac{i}{N}, i \neq i^*$ to time spent at $y^* = \frac{i^*}{N}$. From (14) we compute

$$\rho(i) = \frac{(N-i^*)!i^*!(1 + e^{\mu \Delta r(\frac{i}{N})})e^{-\mu \sum_{j=1}^{i} \Delta r(\frac{j}{N})}}{2(N-i^*)!i^*!e^{-\mu \sum_{j=1}^{i^*} \Delta r(\frac{j}{N})}}. \tag{16}$$

We show that $\rho(i) < 1$ for all $i \notin \{i^* - 1, i^*, i^* + 1\}$ in two steps. In the first case we show that $\rho(i) < 1$ for all $i > i^* + 1$. In the second case we show that $\rho(i) < 1$ for $i < i^* - 1$.

Case 1: Let $\epsilon = i - i^*$ with $\epsilon > 0$. The ratio $\rho(i)$ then becomes

$$\rho(i) = \frac{(N-i^*)!i^*!(1 + e^{\mu \Delta r(\frac{i^*+\epsilon}{N})})e^{-\mu\left(\Delta r(\frac{i^*+1}{N}) + \ldots \Delta r(\frac{i^*+\epsilon}{N})\right)}}{2(N-i^*-\epsilon)!(i^*+\epsilon)!}. \tag{17}$$

Replacing $(N-i)!i!$ in the denominator of (17) with its minimal possible value for $i \in \{0, 1, \ldots, N\}$ yields the inequality

$$\begin{aligned}\rho(i) &\leq \\ &\frac{(N-i^*)!i^*!(1 + e^{\mu \Delta r(\frac{i^*+\epsilon}{N})})e^{-\mu\left(\Delta r(\frac{i^*+1}{N}) + \ldots \Delta r(\frac{i^*+\epsilon}{N})\right)}}{2\lfloor \frac{N}{2}\rfloor!\lceil \frac{N}{2}\rceil!} \\ &\leq \gamma(1 + e^{-\mu \Delta r(\frac{i^*+\epsilon}{N})})e^{-\mu\left(\Delta r(\frac{i^*+1}{N}) + \ldots \Delta r(\frac{i^*+\epsilon-1}{N})\right)}\end{aligned}$$

where $\gamma = \frac{(N-i^*)!i^*!}{2\lfloor \frac{N}{2}\rfloor!\lceil \frac{N}{2}\rceil!}$.

Now assume $\epsilon \geq 2$. Since $\Delta r(\frac{i^*+\epsilon}{N}) > 0$ for all $\epsilon \geq 1$, $\rho(i)$ decreases with increasing ϵ so

$$\begin{aligned}\rho(i) &\leq \gamma(1 + e^{-\mu \Delta r(\frac{i^*+2}{N})})e^{-\mu \Delta r(\frac{i^*+1}{N})} \\ &< \frac{(N-i^*)!i^*!}{\lfloor \frac{N}{2}\rfloor!\lceil \frac{N}{2}\rceil!}e^{-\mu \Delta r(\frac{i^*+1}{N})}.\end{aligned} \tag{18}$$

If (15) is satisfied then (18) becomes $\rho(i) < 1$.

Case 2: Let $\epsilon = i - i^*$ with $\epsilon < 0$. The ratio $\rho(i)$ then becomes

$$\rho(i) = \frac{(N-i^*)!i^*!(1+e^{\mu\Delta r(\frac{i^*-\epsilon}{N})})e^{-\mu\left(-\Delta r(\frac{i^*-\epsilon+1}{N})-\ldots\Delta r(\frac{i^*}{N})\right)}}{2(N-i^*+\epsilon)!(i^*-\epsilon)!} \tag{19}$$

$$\leq \gamma(1+e^{\mu\Delta r(\frac{i^*-\epsilon}{N})})e^{-\mu\left(-\Delta r(\frac{i^*-\epsilon+1}{N})-\ldots\Delta r(\frac{i^*}{N})\right)}. \tag{20}$$

Since $\Delta r(\frac{i^*-\epsilon}{N}) < 0$ for all $\epsilon > 0$ this can also be written

$$\rho(i) \leq \gamma(1+e^{-\mu|\Delta r(\frac{i^*-\epsilon}{N})|})e^{-\mu\left(|\Delta r(\frac{i^*-\epsilon+1}{N})|+\ldots|\Delta r(\frac{i^*}{N})|\right)}. \tag{21}$$

Using the same argument as in Case 1, we arrive at the strict inequality

$$\rho(i) < \frac{(N-i^*)!i^*!}{\lfloor\frac{N}{2}\rfloor!\lceil\frac{N}{2}\rceil!}e^{-\mu|\Delta r(\frac{i^*-1}{N})|}. \tag{22}$$

If (15) is satisfied then (22) becomes $\rho(i) < 1$. □

Theorem 2: If reward structures satisfy Definition 1 and

$$\mu > \mu_2 := \max\left\{-\frac{\ln\left(\frac{2+i^*-N}{N-i^*}\right)}{|\Delta r\left(\frac{i^*+1}{N}\right)|}, -\frac{\ln\left(\frac{2N+2-3i^*}{i^*}\right)}{|\Delta r\left(\frac{i^*-1}{N}\right)|}\right\} \tag{23}$$

then the steady-state choice distribution is maximum for $y = y^*$.

Proof of Theorem 2: Again we examine $\rho(i) = \pi_i/\pi_{i^*}$.
Case 1: Let $\epsilon = i - i^*$ with $\epsilon > 0$.

We assume $\epsilon \geq 1$. We have shown that $\rho(i)$ decreases with increasing ϵ so

$$\begin{aligned}\rho(i) &\leq \frac{N-i^*}{2(i^*+1)}(1+e^{-\mu\Delta r(\frac{i^*+1}{N})})e^{-\mu\Delta r(\frac{i^*+1}{N})} \\ &< \frac{N-i^*}{2(i^*+1)}(1+e^{-\mu\Delta r(\frac{i^*+1}{N})}).\end{aligned} \tag{24}$$

If (23) is satisfied then (24) becomes $\rho(i) < 1$.
Case 2: Let $\epsilon = i - i^*$ with $\epsilon < 0$. We assume $\epsilon = -1$ and (21) becomes

$$\rho(i) \leq \frac{i^*}{2(N-i^*+1)}(1+e^{-\mu|\Delta r(\frac{i^*-1}{N})|}). \tag{25}$$

If (23) is satisfied then (25) becomes $\rho(i) < 1$. □

Example 1: For the matching shoulders reward structure shown in Figure 1, we have $r_A(y) = k_A y + c_A$ and $r_B(y) = k_B y + c_B$ where $k_A = -\frac{1}{2}, c_A = \frac{3}{5}, k_B = 1$ and $c_B = 0$. For this example with $N = 20$, $\mu_1 = 5.11$ and $\mu_2 = 10.81$. These values shrink for smaller N and grow for larger N.
Example 2: For the converging gaussians reward structure shown in Figure 2, we have

$$r_A(y) = e^{-\left(\frac{y-\bar{y}_A}{\sqrt{2}\sigma_A}\right)^2} + c_A, \qquad r_A(y) = e^{-\left(\frac{y-\bar{y}_B}{\sqrt{2}\sigma_B}\right)^2} + c_B$$

with $\bar{y}_A = \frac{2}{5}, \bar{y}_B = \frac{3}{5}$ and $\sigma_A = \sigma_B = \frac{1}{5}$ and $c_A = c_B = \frac{3}{10}$. In this example $\mu_1 = 0$ for any N. For $N = 20$, $\mu_2 = 1.11$ and μ_2 grows almost negligibly with increasing N.

B. Performance Dependence on Model Parameters

Given π, the fraction of time spent at each proportion of choice A, we can compute sensitivity of long-run performance to the parameters of the DDM and task. Here we compute this sensitivity to the parameter μ in the DDM. As mentioned in Section III, larger μ corresponds to increased certainty in the decision making, which can also be interpreted as a reduced tendency to explore.

The average reward can be computed as $\bar{r}(y) = yr_A(y) + (1-y)r_B(y)$. For each value of y, this is the reward that would be received on average if the decision maker were to maintain that value of y. So the expected value of the reward is the sum of each average reward multiplied by the fraction of time spent at each proportion of choice A and is written

$$\tilde{r} = \sum_{i=0}^{N} \pi_i \bar{r}_i. \tag{26}$$

The sensitivity of performance to μ can then be computed as the derivative of the expected value of the reward with respect to μ:

$$\begin{aligned}\frac{d}{d\mu}\tilde{r} &= \sum_{i=0}^{N} \bar{r}_i \frac{d}{d\mu}\pi_i \\ &= \sum_{i=0}^{N}\left(\frac{i}{N}r_A\left(\frac{i}{N}\right) + \frac{N-i}{N}r_B\left(\frac{i}{N}\right)\right)\frac{d}{d\mu}\pi_i.\end{aligned} \tag{27}$$

By denoting $g_i(\mu) := (1+e^{\mu\Delta r(\frac{i}{N})})$ and $M(\mu) := \sum_{j=0}^{N}\pi_j$, the derivative of π_i with respect to μ can be written

$$\begin{aligned}&\frac{d}{d\mu}\pi_i = \\ &\frac{\alpha_i e^{-\mu\beta_i}(\Delta r(\frac{i}{N})e^{\mu\Delta r(\frac{i}{N})} - \beta_i g_i(\mu))}{M(\mu)} - \\ &\frac{\alpha_i e^{-\mu\beta_i} g_i(\mu)\sum_{j=0}^{N}\alpha_j e^{-\mu\beta_j}\left(\Delta r(\frac{j}{N})e^{\mu\Delta r(\frac{j}{N})} - \beta_j g_j(\mu)\right)}{M(\mu)^2}.\end{aligned} \tag{28}$$

Example 1 continued: Consider again the matching shoulders reward structure of Figure 1. The derivative of the expected value of reward with respect to μ, given by (28) is plotted in Figure 3 along with the expected value of the reward. For this reward structure there is a critical point (for $N = 20$ $\mu_c = 1.15$). For $\mu < \mu_c$ increasing μ results in substantially higher reward. However, as μ increases further, the expected value of reward decreases. This is an example for which some exploratory behavior in the decision making is beneficial and is directly related to the results of Theorems 1 and 2. For instance, in the case $N = 20$, for $\mu > \mu_1 = 5.11$ there is not a lot of exploratory behavior and the decision maker converges to the matching point of the reward structures, which is not the optimal strategy.

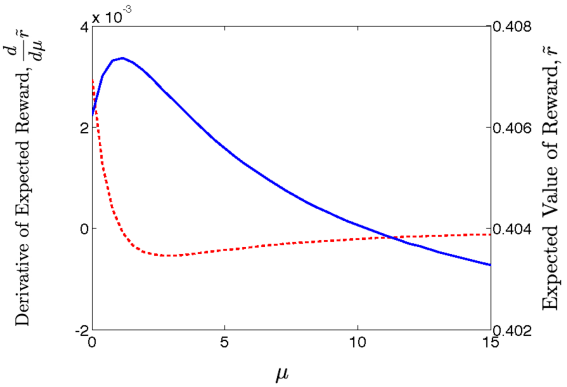

Fig. 3. The derivative of the expected value of reward for the matching shoulders reward structure shown in Figure 1. The dotted line is $\frac{d}{d\mu}\tilde{r}$ from Equation (28). The solid line is the expected value of reward, $\tilde{r}$. Both are plotted against μ for $N = 20$.

Example 2 continued: Consider again the converging gaussians reward structure of Figure 2. The derivative of the expected value of reward with respect to μ, given by (28) is plotted in Figure 4 along with the expected value of the reward. In this example, $\frac{d}{d\mu}\tilde{r}$ is positive for all μ.

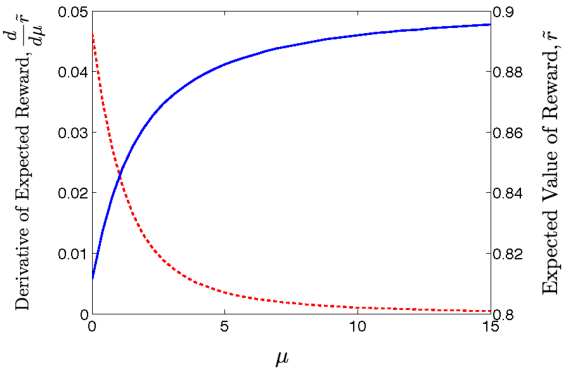

Fig. 4. The derivative of the expected value of reward for the converging gaussians reward structure shown in Figure 2. The dotted line is $\frac{d}{d\mu}\tilde{r}$ from Equation (28). The solid line is the expected value of reward, $\tilde{r}$. Both are plotted against μ for $N = 20$.

The derivative is always positive in this example (for any N) because the matching point coincides with the maximum of the expected value of reward; i.e., when the decision maker converges to y^* in the converging gaussians reward structure, it is also the case that the highest reward on average is received. Therefore, increasing the parameter μ, or the certainty in the decision making, results in higher expected reward for the task. We note, however, that there is not a great deal of gain in performance once μ increases above a threshold approximately equal to 5.

VII. Final Remarks

In this paper we prove conditions for which the DDM for the TAFC task is a Markov process. This allows us to derive transition probabilities and analytical expressions for steady-state distributions of choice sequences as a function of DDM and TAFC task parameters. We use the expressions to prove results about the long-run decision dynamics. In particular we prove conditions on DDM parameter μ, which measures the level of certainty or tendency to explore in the decision maker, that lead to matching behavior. We also study performance sensitivity to the parameter μ. We apply the results to two example reward structures.

In ongoing work, motivated by the investigations in [8], we are extending our modeling and analysis approach to address multiple human decision makers, engaged in TAFC tasks, that exchange information on their choices or performance. In this case there is a DDM for each decision maker and the models are coupled by feedback between individuals.

References

[1] B. Donmez, M. L. Cummings, and H. D. Graham. Auditory decision aiding in supervisory control of multiple unmanned aerial vehicles. *Human Factors: J. Human Factors and Ergonomics*, In press.

[2] K. C. Campbell, Jr. W. W. Cooper, D. P. Greenbaum, and L. A. Wojcik. Modeling distributed human decision-making in traffic flow management operations. In *3rd USA/Europe Air Traffic Management R&D Seminar*, Napoli, June 2000.

[3] A. R. Otto, T. M. Gureckis, A. B. Markman, and B. C. Love. Navigating through abstract decision spaces: Evaluating the role of state generalization in a dynamic decision-making task. *Psychonomic Bulletin and Review*, In press.

[4] P. R. Montague and G. S. Berns. Neural economics and the biological substrates of valuation. *Neuron*, 36:265–284, 2002.

[5] T.M. Gureckis and B.C. Love. Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology*, 53:180–193, 2008.

[6] D. M. Egelman, C. Person, and P. R. Montague. A computational role for dopamine delivery in human decision-making. *Journal of Cognitive Neuroscience*, 10:623–630, 1998.

[7] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113:700–765, 2006.

[8] A. Nedic, D. Tomlin, P. Holmes, D.A. Prentice, and J.D. Cohen. A simple decision task in a social context: experiments, a model, and preliminary analyses of behavioral data. In *Proc. of the 47th IEEE Conference on Decision and Control*, 2008.

[9] B. K. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Springer-Verlag, Berlin, 2003.

[10] P. Simen and J. D. Cohen. Explicit melioration by a neural diffusion model. 2008. Submitted to *Brain Research*.

[11] R. Bogacz, S. M. McClure, J. Li, J. D. Cohen, and P. R. Montague. Short-term memory traces for action bias in human reinforcement learning. *Brain Research*, 1153:111–121, 2007.

[12] P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, 16:1936–1947, 1996.

[13] R. S. Sutton and A. G. Barto. *Reinforcement learning*. MIT Press, Cambridge, MA, 1998.

[14] H. M. Taylor and S. Karlin. *An Introduction to Stochastic Modeling -3rd ed.* Academic Press, 1998.

[15] R. Herrnstein. Rational choice theory: necessary but not sufficient. *American Psychologist*, 45:356–367, 1990.

[16] R. Herrnstein. *The Matching Law: Papers in Psychology and Economics*. Harvard University Press, Cambridge, MA, USA, 1997. Edited by Howard Rachlin and David I. Laibson.

[17] M. Cao, A. Stewart, and N. E. Leonard. Integrating human and robot decision-making dynamics with feedback: Models and convergence analysis. In *Proc. 47th IEEE Conf. on Decision and Control*, 2008.

[18] M. Cao, A. Stewart, and N. E. Leonard. Convergence in human decision-making dynamics. *Systems and Control Letters*, 59:87–97, 2010.

[19] L. Vu and K. Morgansen. Modeling and analysis of dynamic decision making in sequential two-choice tasks. In *Proc. 47th IEEE Conf. on Decision and Control*, 2008.